Poster I-16

Combining Extracted Gene Relations With Ontologies To Create Meaningful Pathway Maps

***McDonald, Daniel*[*1]*, Chen, Hsinchun*[*1]*, Su, Hua*[1]*, Marshall, Byron*[1]*, Zhang, Yiwen*[1]*, Eggers, Shauna*[1]*, Tseng, Chun-Ju*[1]*, Martinez, Jesse*[2]**
[1]Artificial Intelligence Lab, University of Arizona, Tucson, AZ, USA; [2]Arizona Cancer Center, University of Arizona, Tucson, AZ, USA

Today, biomedical research findings outpace the ability of any one researcher to stay on top of all the literature. The MEDLINE database is a valuable source of information for such research. The collection contains information for over 12 million articles and continues to grow at a rate of 2,000 articles per week. Using natural language processing techniques, we extract open-ended genetic regulatory relationships from the MEDLINE abstracts using two different parsers. The first parser uses preposition-based templates to extract information triplets from the text, primarily in the form of noun relations. The second parser identifies grammatical verb relations and extracts only those relations that appear in certain ontologies. We have evaluated the use of the Gene Ontology (GO), the approved gene names from the Human Genome Organization (HUGO), and the UMLS Specialist Lexicon. Three findings resulted. First, integrating ontologies with the extraction process produced more relations that could be linked to ontologies for further aggregation and disambiguation. Second, using ontologies increased the precision and therefore quality of the extracted relations. Third, the use of different ontologies created different themes in the relations extracted. Using 42 abstracts, an expert categorized the correct relations extracted into one of three categories: a pathway relation, a biologically relevant relation, or an acceptable, but incomplete relation. Results indicated that using the combination of GO and HUGO produced a larger percentage of pathway relations while using the UMLS Specialist Lexicon produced a larger percentage of biologically relevant relations. Such results indicate a complementary role for different ontologies when extracting relations.

Now that gene relations have been extracted from the text and connected to ontologies, we are developing a model for aggregating the text relationships, and presenting them to biomedical researchers in an intuitive manner. As part of the aggregation process, we establish links between relations based on relationships shared by their ontological concepts. During aggregation, we identify contradictions in the literature and measure the strength of relations based on the repetition of the findings in the literature as well as the rating of the publishing journal. We have built three collections of relations focusing on P53, AP1, and Yeast. A visual exploration tool has been developed using a spring algorithm to show gene interactions. The gene network is dynamically expanded when nodes are selected. Automatically generating gene pathway networks using relations extracted from literature and arranged via ontologies can help medical researchers manage their literature reviews and possibly draw conclusions that would not have been possible otherwise.

This research was sponsored by the following grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: a Toolkit for Gene Pathway Analysis"